USING THE LITERATURE IN THE STUDY OF EMERGING FIELDS OF SCIENCE AND TECHNOLOGY

Michael A. Rappa
Massachusetts Institute of Technology

Raghu Garud
New York University

July 1991

Sloan WP # 3392-92

Massachusetts Institute of Technology

USING THE LITERATURE IN THE STUDY OF EMERGING FIELDS OF SCIENCE AND TECHNOLOGY

Michael A. Rappa
Massachusetts Institute of Technology

Raghu Garud
New York University

July 1991

Sloan WP # 3392–92

Alfred P. Sloan School of Management
Massachusetts Institute of Technology
50 Memorial Drive, E52-538
Cambridge, MA 02139-4307

USING THE LITERATURE IN THE STUDY OF EMERGING FIELDS OF SCIENCE AND TECHNOLOGY: AN EXAMINATION OF RESEARCHER CONTRIBUTION-SPANS

Michael A. RAPPA
Massachusetts Institute of Technology

Raghu GARUD†
New York University

ABSTRACT

This study describes how the scientific and technical literature can be used as a source of data to analyze the emergence of new fields. In particular, the literature is used to determine the length of contribution spans of individual researchers in a field. Through the use of statistical techniques, such as survival analysis, the authors examine the probability that a researcher will contribute to the field for a specified length of time and the probability that a researcher, having contributed to the field for a specified period of time, will cease to contribute in the future. The authors also test the significance of several variables in explaining researchers' contribution spans.

INTRODUCTION

New fields of science and technology typically emerge in the context of a community of individuals who develop a common set of ideas and techniques and who communicate with each other about the nature of their work. Frequently, a portion of the communication between these individuals takes the form of conference presentations and scientific papers, which then become documented in the scientific and technical literature. When used to its fullest and in conjunction with other kinds of data, this information can be extremely helpful in improving our understanding of the dynamics of a research community.

The following study, which serves as one illustration of this approach, uses data from the literature to measure the length of time researchers remain in a field, thereby determining statistically their survival and hazard rates as well as some of the factors associated with their longevity. In particular, it examines the duration of the participation of individual researchers in a field through an analysis of their "contribution-spans": that is, the time span between their first and last contribution to the published literature in a given field. From an analysis of the contribution-spans of researchers in the field of cochlear implants, estimates are made of: (1) the probability that a researcher's contribution-span will extend a given number of

---

† Michael Rappa is assistant professor of management with the Massachusetts Institute of Technology. Raghu Garud is assistant professor of management with New York University. The authors are indebted to Susan Bjørner for her assistance in conducting the literature searches and Gary Quick for his assistance in creating the database. The authors also thank Joel Baum, Hayagreeva Rao and Robert Yaffee for helpful comments and guidance, and Edward Roberts for his support and encouragement. This study was funded, in part, with a grant from the Council on Library Resources.

years, and (2) the probability that, having contributed a given number of years, a researcher will cease to contribute in the future. Furthermore, this paper examines the relevance of a number of individual-, organizational-, and community-related variables in explaining the duration of researchers' contribution-spans.

Before elaborating upon the concept of contribution-spans and discussing the procedures and results of this study, a brief review of earlier studies that have used the literature in assessing the growth of new fields is in order.

Prior Literature-based Studies of Emerging Fields

The use of the literature to understand the growth of a field has a long and well-established tradition. When Cole and Eales published their study of the development of comparative anatomy in 1917, they were among the first to seriously utilize the published literature in order to quantify the progress in a field.[1] Given the technology of the day, it was truly a painstaking effort by the two scientists, who examined nearly 6,500 books and papers to compile their data. The result was a detailed statistical account of the ebb and flow of research in comparative anatomy over three centuries. Cole and Eales clearly illustrated the prevalence and magnitude of cyclical changes in the level of publication activity that occur in a field over time. This finding led them to suggest that such waves of activity can be attributed, in part, to the movement of individuals between different research fields.

The pioneering effort of Cole and Eales was joined subsequently by Wilson and Fred, who published a study in 1935 of the growth and development of research on the subject of nitrogen fixation by plants.[2] It is clear that Wilson and Fred had an intuitive grasp of the value of the literature as something more than a means of communicating research results among scientists. "Study of the literature as an entity...," they claim, "...should provide valuable information for the interpretation of past production and might afford some basis for prediction of future trends" within a given field.[3] Like their predecessors, Wilson and Fred showed empirical evidence of the cyclical nature of research within a field. Similarly, they suggest that the fluctuating level of publication activity in nitrogen fixation research was

---

[1] F.J. Cole and N.B. Eales, "The History of Comparative Anatomy, Part I.—A Statistical Analysis of the Literature," *Science Progress* 11 (1917): 578-96.

[2] P.W. Wilson and E.B. Fred, "The Growth Curve of a Scientific Literature: Nitrogen Fixation by Plants," *Scientific Monthly* 41 (1935): 240-50.

[3] Ibid., p. 240.

a result of the movement of researchers between fields. They claim that the level of activity would rise with an important breakthrough as it attracted new scientists to the field and led to further experiments; and afterward it would subside as scientists found the remaining problems too complicated to solve.

It was not for another two decades that the study of the literature would once again receive serious attention. The one most responsible for this resurgence was Price, who in 1956 used the physics literature to document the "exponential growth of science."[4] Price's work brought legitimacy to the scholarly examination of literature, partly because his conclusions coincided with, and indeed reinforced, the belief that the proliferation of science seemingly knew no bounds. More recently, literature-based studies of emerging fields of science and technology have become fairly common. The nature of this research is quite diverse. However, in general, there have emerged two basic types of studies. The first seeks to model the growth of a field by measuring annual publication volume.[5] The second is quite different in that it uses the citation as a unit of analysis.[6] Citation-based studies seek to understand the interlocking nature of citation patterns as a means for observing the development of clusters of researchers who may ultimately come to form the basis for a new field. Both types of literature studies cover a wide range of topics in science and technology, although attention to the latter has recently led to a more intensive examination of the patent literature.[7]

The contributions of Cole and Eales, Wilson and Fred, and Price, among many others, have established a tradition of research that uses the literature to assess the growth and direction of a field. However, in having led the way in revealing the promise of using the literature, their work has exposed some of its limitations, as well. Perhaps the most

---

[4]D.J. Price, "The Exponential Curve of Science," *Discovery* 17 (1956): 240-43.

[5]For example, see A. Granberg, "A Bibliometric Survey of Fiber-Optics Research in Sweden, West Germany, and Japan," Monograph, Research Policy Institute, University of Lund, Sweden, 1985.

[6]The notion of "bibliographic coupling" was proposed by Kessler in 1962. Since then, numerous "co-citation" studies have been published, including the pioneering work of Small and his colleagues. For a recent example, see H. Small and E. Greenlee, "Collagen Research in the 1970s," *Scientometrics* 10 (1986): 95-117. A variant of co-citation analysis is co-word analysis (alone or in combination with co-citation) sometimes referred to as "science mapping." See P. Healy, H. Rothman, and P.K. Hoch, "An Experiment in Science Mapping for Research Planning," *Research Policy* 15 (1986): 233-51; R.R. Brahm, H.F. Moed, and A.F.J. van Raan, "Mapping of Science: Critical Elaboration and New Approaches," *Informetrics* (1987/88): 15-28; and M. Callon, J. Law and A. Rip (eds.) *Mapping the Dynamics of Science and Technology* (London: Macmillian Press, 1986).

[7]For an overview, see B.L. Basberg, "Patents and the Measurement of Technological Change: A Survey of the Literature," *Research Policy* 16 (1987): 131-41. An early study of patents is provided by W.D. Reekie, "Patent Data as a Guide to Industrial Activity," *Research Policy* 2 (1973): 246-64; for a more recent study, see R.M. Wilson, "Patent Analysis using Online Databases: Technological Trend Analysis," *World Patent Information* 9 (1987): 18-26.

troublesome criticism of prior studies is that the literature is a singular, coarse-grained indicator of the level of research activity in a field. While it may be true that the literature offers a useful measure of scientific and technological output, the focus on publication and patent statistics alone tells us little else about the process of the emergence of a new field.

Although the criticisms are valid, it would be unfortunate if they were to dampen interest in the use of the literature for the study of new fields of science and technology. The literature is a rich source of information. Many of the limitations of previous studies are not necessarily inherent in the data, but rather they arise from the inability of these studies to exploit the full potential of the literature. However, this situation is changing. Recent advances in computer hardware and software are making it possible to use the literature with greater ease and sophistication, enabling investigators to go beyond measures of publication volume and citations to take full advantage of the information it contains.

Once the perspective is shifted from treating the literature as something to be measured in and of itself, to using the literature *as a source of data* about what goes on in research communities, many new avenues of research become readily apparent. In this vein, a small number of scholars have sought to use the literature as a source of data about research communities and the emergence of new fields. One pioneering effort is Mullins' 1972 study of molecular biology, in which he takes advantage of the incidence of co-authorship in the literature to understand the communication network that forms among researchers in the field.[8] Taking a different approach, Comroe and Dripps show how the content of the literature can be analyzed to understand the contribution of long-term basic research to major advances in clinical medicine.[9] Yet another approach is that of Spiegel-Røsing, who uses the literature to identify individual researchers in order to compare the level of scientific manpower in different countries.[10] Notice that in each of these instances, the investigators went beyond numbers of publications and instead probed the literature for specific

---

[8]N.C. Mullins, "The Development of a Scientific Specialty: The Phage Group and the Origins of Molecular Biology," *Minerva* 10 (1972): 51-82. Since Mullins, others have sought to use co-authorship data to investigate collaboration among researchers, including constructing a map of the inter-organizational linkages within a community. For example, see D. deB. Beaver and R. Rosen, "Studies in Scientific Collaboration," *Scientometrics* 1 (1978): 65-84; M.S. Sridhar, "A Study of Co-authorship and Collaborative Research Among Indian Space Scientists," *R&D Management* 15 (1985): 243-49; and M.A. Rappa, "Assessing the Emergence of New Technologies: The Case of Compound Semiconductors," in *Research on the Management of Innovation*, A. Van de Ven et al. (Cambridge, Mass.: Ballinger, 1989).

[9]J.H. Comroe, Jr. and R.D. Dripps, "Scientific Basis for the Support of Biomedical Science," *Science* 192 (1976): 105-111.

[10]I.S. Spiegel-Røsing, "Journal Authors as an Indicator of Scientific Manpower: A Methodological Study Using Data for the Two Germanies and Europe," *Science Studies* 2 (1972): 337-59.

information contained therein: that is, about the people involved, their interaction with each other, or the nature of their work.

These studies have only just begun to tap the abundance of information that is contained within the literature. Journal articles, conference papers, and patents in a given field represent a detailed, self-reported archival record of the effort generated by researchers to solve the scientific and technical problems confronting them. Furthermore, the literature is an appealing source of data in several respects: the conventions of publication ensure a level of quality and authenticity; the data can be collected unobtrusively; the findings can be replicated and tested for reliability; and the data are publicly available and relatively inexpensive to collect. Clearly, it would be very difficult to match the comprehensive scope and longitudinal nature of the literature using other data collection techniques. When taken together, the literature can be viewed as a unique chronology of the efforts of researchers to establish a new field, and can provide information with respect to the individuals involved, where they are employed, who they collaborate with, what problems they are pursuing, and when they were active in the field. This paper illustrates how such data can be used to obtain a better understanding of the longevity of researchers' contributions during the emergence of a new field.

Researcher Contribution-spans

The challenge to understanding the emergence of new fields of science and technology is, in part, a problem of understanding why researchers choose the topic they do, and why they remain committed to that topic or leave it for something else. The fundamental assumption is that fields which attract and retain researchers are likely to progress more quickly than those that are less attractive and are therefore less able to recruit and retain researchers. Furthermore, it is reasonable to assume that researchers' decisions to remain with a field are influenced by their assessment of the rate of progress being made. Given this behavior, it may be interesting to use the literature to understand in a statistical manner the duration of researchers' participation in a given field and the factors that may influence it.

Determining the number of years a researcher spends working in a field is not a simple task. First, it requires some degree of historical investigation: if we simply ask everyone who currently works in a particular field how long they have participated, we would overlook the experience of numerous individuals who have already come and gone over the years. Second, the number of individuals who have participated in a field can be quite large and can be widely dispersed geographically, thereby making it very hard to ascertain the extent of their

participation. Confronted with these constraints, it becomes appealing to look to the literature as a source of data about who participates in a field and for how long. Thus, the "contribution-span"—that is the number of years spanning an author's first and last known publications in a field—can serve as a unique and useful measure of the duration of their participation in a field.

Using the statistical techniques of survival analysis, data on researcher contribution-spans derived from the literature can be used to estimate the probability that a researcher will remain in the community a given number of years (the survival rate). Furthermore, the data can also be used to estimate the conditional probability of a researcher leaving the field after contributing a given number of years (the hazard rate). However, once the distribution of the survivor and hazard functions is understood, it becomes pertinent to ask what factors might influence researcher contribution-spans. Here again, the literature can offer some insights in terms of providing additional data regarding each researcher. For example, the length of an individual's contribution-span might be affected by: (1) their own or their organization's cumulative productivity in the field; (2) the size of their collaborative social network; (3) the size of their organization's research effort in the field; (4) the kind of organization in which they are employed and its geographic location; (5) the accumulated amount of knowledge in the field; (6) the number of other researchers working in the field and the extent of their dispersion among different organizations; and (7) the evolutionary stage of development of the field. Using data from the literature, the statistical relationship between each of these factors and researchers' contribution-spans will be examined.

The Cochlear Implants Field

The field of cochlear implants was chosen as an illustrative case for the present analysis, but it is just one of several fields that the authors are studying as part of a larger research program on the assessment of emerging areas of science and technology.[11] Although studies of the electrical stimulation of the ear have a long history, it was not until the 1950s that researchers began the systematic investigation of how electrical stimulation might provide a means for enabling individuals with sensorineural deafness to gain some sense of hearing. One result of this research was the advent of the cochlear implant, a device that uses electrical stimulation of the cochlear to provide a sense of sound for profoundly deaf individuals.[12]

---

[11] For a full account of the historical development of cochlear implants, see R. Garud and A.H. Van de Ven, "Technological Innovation and Industry Emergence: The Case of Cochlear Implants," in Van de Ven, et al., *Research on the Management of Innovation* (Cambridge, Mass.: Ballinger, 1989).

[12] A cochlear implant is composed of a microphone, signal processor, and transmitter worn on the outside of the body, and a receiver that is surgically implanted behind the ear just under the skin either into the cochlear or placed around it. For a

One of the first trials of a rudimentary cochlear implant device was performed in 1961 by William House, a clinical physician, who later founded the House Ear Institute, a major center for cochlear implant research. At first, few researchers took an interest in cochlear implants, largely due to the intense controversy the field encountered. Many considered the device to lack a scientific basis and believed that it was too premature for human experimentation. It was not until the early 1970s that the controversy subsided and cochlear implants emerged as a viable research field. A flurry of meetings were held in 1973 that galvanized the burgeoning field: including, a workshop held as part of the American Otological Society meeting (a group that was adamantly opposed to the topic in previous years), a major international conference on the subject and a workshop at the University of California, San Francisco. As a result of these meetings, a limited number of clinical trials began.[13]

One study of particular importance is Bilger's work, initiated in 1975, on the comparative performance of cochlear implants.[14] Funded by the U.S. National Institutes of Health and released in 1977, the "Bilger Report" showed that although the early claims regarding performance were exaggerated, cochlear implants did indeed have merit. In its conclusions, the report lent some sorely needed legitimacy to the field and thereby opened the door for a greater number of researchers to participate. This growth in participation can be seen clearly in terms of the number of researchers contributing to the literature (See Figure 1).[15]

It is estimated that by 1990 there were about four-hundred individuals active in the field, who were employed in nearly one-hundred different organizations worldwide. About forty-percent of the cochlear implant community is located in the U.S. Although tremendous progress has been made over the years and about three-thousand patients have received the

---

detailed description of cochlear implants see: Gerald E. Loeb, "The Functional Replacement of the Ear" *Scientific American*, 225 (2):104-11.

[13]Proceedings of the First International Conference on the Electrical Stimulation of the Acoustic Nerve as a Treatment for Profound Sensorineural Deafness in Man, San Francisco, California, 1974 (edited by M.M. Merzenich, R.A. Schindler and F. Sooy); Report on a Workshop on Cochlear Implants held at the University of California at San Francisco, October 23-25, 1974 (edited by M.M. Merzenich and F. Sooy).

[14]R.C. Bilger, et al., "Evaluation of Subjects Presently Fitted With Implanted Auditory Prostheses," *Ann. Otol. Rhinol. Laryngol.*, (Suppl. 38) 86 (1977): 3-10.

[15]The number of researchers in the community in a given year is calculated to be the cumulative number researchers entering the field (as evidenced by an initial publication) subtracted by the cumulative number of individuals who have left the field (as evidenced by their failure to continue to publish in a future year).

cochlear devices, the cochlear implant remains largely an experimental procedure with many challenging problems to be overcome.[16]

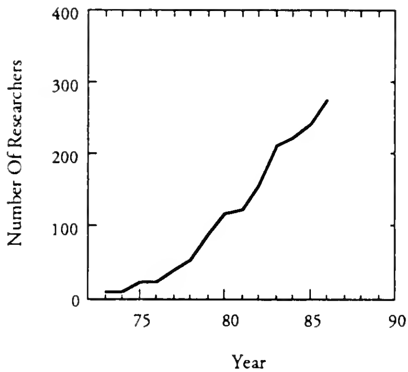

FIGURE 1: *Growth of Cochlear Implant Research Community, 1973-89*

DATA COLLECTION AND METHODS

Five commercial electronic databases covering the medical science and engineering literature were used to identify publications related to the field of cochlear implants.[17] The databases were searched on-line using a set of key terms that are known to be commonly used in the lexicon of cochlear implant researchers and might be either in the title, abstract or classification terms of a document. The search strategy was derived from an earlier one used by the National Library of Medicine to create a research bibliography on cochlear implants.[18]

---

[16]Several researchers have provide historical accounts of the field of cochlear implants and its technical evolution. For example, see W.F. House and K.I. Berliner, "Cochlear Implants: From Idea to Clinical Practice," in H. Cooper, ed., *Practical Aspects of Cochlear Implants* (London: Taylor and Francis) in press; and F.B. Simmons, "History of Cochlear Implants in the United States," in R.A. Schindler and M.M. Merzenich, eds. *Cochlear Implants* (New York: Raven Press) 1985.

[17]The databases used are Medline, Compendex, Biosis, Exerpta Medica and INSPEC. A total of 3064 documents were originally identified as related to cochlear implants. After implementing an on-line duplication removal process, the database was reduced to 1884 documents.

[18]K. Patrias and R.F. Naunton, National Library of Medicine, *Current Bibliographies in Medicine: Cochlear Implants,* January 1983-March 1988 (Washington, DC: U.S. Department of Health and Human Services, National Public Health Service, National Institute of Health, 1988).

The cochlear implant documents retrieved electronically from the search were temporarily placed in a bibliographic relational database operating on a personal computer. This allowed for a careful inspection of each document in order to ensure the accuracy and integrity of the search procedure. Since multiple source databases were used, it was necessary to remove duplicate documents. In addition, while inspecting the database, an effort was made to remove misclassified documents that did not pertain to cochlear implants as well as those that did not have a research orientation, such as editorials or journalist documents. In the process of inspecting the documents, any that seemed inappropriate were flagged, so that an individual active in cochlear implant research could make the final judgment as to its relevance.

The data collection procedure described above ultimately resulted in the identification of 1,329 unique documents related to cochlear implants published between 1973 and 1989.[19] The data subsequently used in this study were derived from these documents. However, before the documents could be used as a source of data, they required extensive editing in order to create a consistency among author names and affiliation names. It is frequently the case that the name of an author or an affiliation is not standardized across documents, especially when using different commercial databases. Sometimes this arises because of misspellings, but mostly this is the result of variations in the use of abbreviations, middle initials, capitalizations, hyphenations, and other sources of inconsistency.[20] Although such a lack of standardization might not be a problem for the typical user of an electronic literature database, it would be a major source of error in determining the duration of researcher contribution-spans. Therefore, it was essential to meticulously inspect the name of each author and affiliation in the relational database so that all inconsistencies could be eliminated.

Upon completing the editing of the documents, the database was used to identify each individual author who contributed to the field over the seventeen-year period. This procedure yielded a total of 1,257 authors. At this stage, a statistical database was created containing several individual-, organizational-, and community-level variables for each

---

[19]Note that the electronic databases do not provide any indication of the work in cochlear implants prior to 1973. This is due in part to the fact that many electronic literature database services did not begin until the late sixties and early seventies. It also is due to the lack of publication of the pre-seventies work in cochlear implants in journals abstracted by the database services. Unfortunately, this is one limitation of the electronic databases that is extremely difficult and costly to circumvent.

[20]The prevalence of author name inconsistencies is examined by M.L. Pao, "Importance of Quality Data for Bibliometric Research," *National On-Line Meeting (10th) Proceedings*, May 9-11, 1989.

author that were derived from information obtained from the published documents. Table 1 provides a list of the variables and their definitions.

The dependent variable for the analysis, the contribution-span, is calculated as the number of years that have elapsed from the first to the last known publication for each author.[21] Although calculating the contribution-span is relatively straightforward, there are some methodological issues that arise that require further explanation. The primary issue of concern is that for those researchers who are active in the field of cochlear implants at the present time, the ultimate length of their contribution-span is indeterminate: that is, since these individuals have not yet left the field, it is only known that the length of their contribution-span is some minimum value (that is, the entry year to the present year). To account for this, survival analysis statistics were implemented in analyzing the data.[22] Such techniques take into consideration precisely this kind of problem in the calculations with a procedure that adjusts for the biases that right-censored data create.

Determining whether or not a researcher is still active in the field can be difficult in certain cases. The reason for this is that typically researchers do not publish every year, and indeed, an author's contribution-span in a field can be characterized by "gaps" of several years in duration in which there are no publications to their credit. The existence of discontinuities in publication records raise the issue of how frequently a researcher must publish in order to be considered an active contributor to the field. Thus, understanding the nature and prevalence of gaps in a researcher's contribution-span is important in determining the proper censoring scheme to use in the analysis. The question arises: How long after someone ceases to publish is it reasonable to assume they are no longer in the field? The answer to this question is necessary in order to determine who has exited the field and who continues to be a participant.

Normally, the time between publications may be a year or two, but in some instances it can be quite long. Therefore, a rule is required to determine how many years should transpire after the last publication in order for it to be reasonable to classify a researcher as having exited the field. An analysis of the frequency of gaps between publication in researcher contribution-spans provides the evidence on which to base this decision (see Figure 2). The

---

[21]For example, if a researcher first published in 1975 and last published in 1980, the researcher's contribution span would be calculated as six years. Furthermore, it is assumed that a researcher who publishes in only one year has a span of one year. Note that the contribution span is unaffected by the frequency of publication within a given year.

[22]See R.C. Elandt-Johnson and N.L. Johnson, *Survival Models and Data Analysis* (New York: John Wiley & Sons, 1980) and J.D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data* (New York: John Wiley & Sons, 1980).

cochlear implant data indicate that 242 (19%) researchers have a gap in their contribution-spans. Among those who have a gap in their contribution-span, four out of five researchers have a gap of three years or less in duration. Based upon this evidence, it was decided that researchers who published within the past three years of the last year of the data set (1989) would be censored.

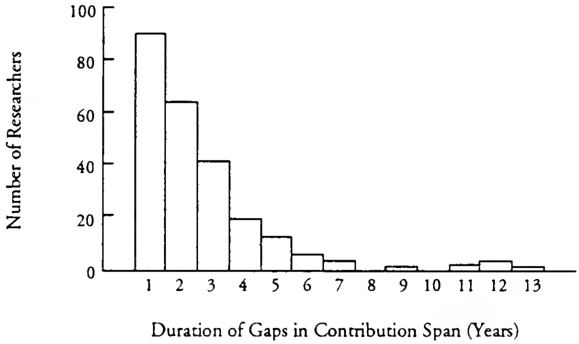

FIGURE 2: *Frequency of Gaps Between Publication in Researcher Contribution-spans*

Only a small number (3.7%) of all researchers have a gap between publications of longer than three years in duration, and still fewer have an exceptionally large gap, such as ten years or more. Although rare, these "sparse" contribution-spans may be indicative of an individual who does not in fact continuously contribute to the field; and thus, although their contribution-span might be quite long, their actual participation in the field is far less than the span implies. Because of their relative rarity in the present data, researchers with sparse contribution-spans were not treated as special cases in the analysis. Nevertheless, it is advisable to address this issue more closely in future studies.

Having determined the distribution of contribution-spans, it is interesting to examine what factors might affect how long a researcher contributes to the field. Using the literature, a number of explanatory variables were constructed. Although the explanatory variables could be treated as time-varying covariates (that is, as having values that vary yearly in the course of an author's contribution-span), the present analysis implements a "single-spell"

approach to formulating the data set.[23] Therefore, the value for each explanatory variable is taken according to the last year in the author's contribution-span. In this manner, several variables were created to control for certain factors that might account for heterogeneity among researchers. First, the kind of organization in which each author is employed was coded according to whether they reside in an academic, industrial, government, or independent research laboratory.[24] Second, the country in which the author is located was coded. Since nearly half of the authors were based in the United States, to simplify the analysis, a location variable was coded as to whether the author resided in the U.S. or not. Third, a variable was created to determine whether the author left the field during the "pioneering" phase (considered to be prior to the publication of the Bilger Report), or during the "rapid growth" phase that occurred from 1978 onwards. The last control variable is a measure of the accumulated knowledge in the field, as determined by the cumulative number of publications.

Additional explanatory variables were created, which are grouped according to their level of analysis: individual-, organizational-, and community-level. At the individual-level, a variable was constructed to reflect an author's cumulative productivity in the field as measured by the cumulative number of publications to their credit. In addition, a variable was created to reflect the extent to which an author is embedded in a larger social network of collaborators. This network is measured by the cumulative number of unique individuals with which an author has been associated with as a co-author on publications.

Two organizational-level variables were created to reflect the size and productivity of an author's institutional affiliation. The size of an author's organization is measured in terms of the number of individuals affiliated with that organization who also publish in the field. Cumulative organizational productivity is measured as the cumulative number of cochlear implant publications by individuals affiliated with an author's organization.

Three community-level variables were created to reflect the size and dispersion of the cochlear implant field in each year. Population size is measured in terms of the number of individual authors who publish in the field in a given year. A second-order variable, the square of population size, was created in order to capture any quadratic association between

---

[23]An analysis that implements time-varying covariates would require a multiple-spell data structure. This work is currently being performed and the results will be described by the authors in a future report.

[24]It is important to note that one unfortunate limitation of electronic literature databases is the tendency of providing affiliation data for the lead author only. Therefore, in some instances secondary authors who are affiliated with a different organization may be misclassified.

population size and contribution-span. The third variable is a measure of dispersion of authors among different organizations: that is, the extent to which the community is concentrated in a few organizations or spread across many. For this purpose, a Hirfindahl statistic, which is determined by calculating the sum of the squared share of researchers affiliated with each organization, in used.

RESULTS

Using data from the scientific and technical literature on cochlear implants published between 1973 and 1989, the contribution-spans for 1,257 researchers and several explanatory variables associated with each were compiled into a statistical database. The data were analyzed using the LIFETEST and LIFEREG procedures of SAS (version 5.18). Of the 1,257 cases, 700 (55.7%) were active within three years of the last year of the data, and were therefore classified as censored.

Using the LIFETEST procedure, the first step in the analysis was to make non-parametric estimates of the survival and hazard functions for the data. The lifetable approach was chosen. The results of this procedure are illustrated in Figure 3. The survival function is negatively-sloped and non-monotonic. The median survival time in 5 years: that is, half the sample leave the field within five years on their first publication. The probability of a researcher's contributions-span lasting longer than two years is about 0.6. The survival rate continues to diminish rapidly between two and six years, and then levels-off eventually reaching a value of about 0.33 for contribution-spans of ten years or more.

The hazard rate is also a negatively-sloped non-monotonic function. The hazard rate decreases very rapidly for researchers who have contribution-spans of at least two years: that is, the probability of a researcher ceasing to contribute after having contributed for two years is only about 0.08 compared to about 0.25 for a researcher in the field only one year. The hazard rate stabilizes for researchers whose contribution-spans are between two and six years, and than diminishes rapidly between six and twelve years. The basic implication of the hazard function is that the longer a researcher contributes to the field, the less likely he or she is to leave it, with the first and sixth years being particularly critical points. Indeed, the risk of leaving the field is highest within the first year.

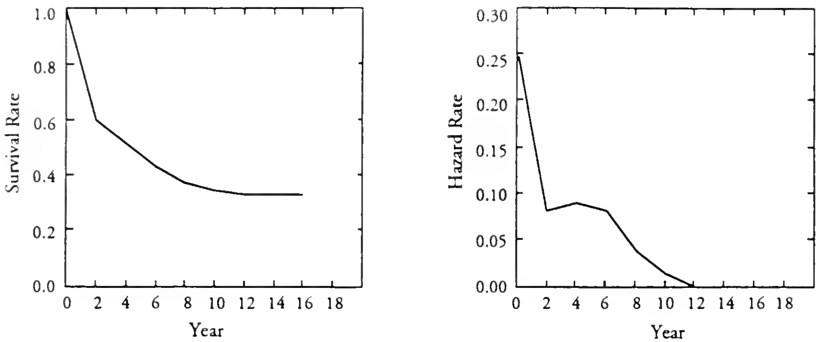

FIGURE 3: *Non-parametric estimation of survival and hazard functions for researcher contribution-spans in cochlear implants.*

The next step in the analysis was to determine the parametric model that best fits the distribution of contribution-spans (or, in the terminology of survival analysis, the failure time). Although non-parametric analysis permits certain assumptions that can be made about the shape of the failure distribution (for instance, that it is non-monotonic), nonetheless it was decided to statistically examine several different distributions for goodness of fit. The basic model adopted for the analysis is:

$$Y = X\beta + \sigma\varepsilon$$

where Y is the log of the contribution-span (the failure time), X is the matrix of explanatory variables, β is a vector of unknown regression parameters, σ is a scale parameter and ε is a vector of errors from an assumed distribution. This model is often referred to as an accelerated failure time model because the effect of the explanatory variables is to scale a baseline distribution of failure times. Specifically, four different types of distributions were evaluated: the exponential, Weibull, gamma, and log-logistic distributions. The results of this procedure are provided in Table 2. The parameters are estimated by maximum likelihood using a Newton-Raphson algorithm. The overall fit of each model is represented by the log-likelihood function. Minus two times the log-likelihood value has a χ^2 distribution with appropriate degrees of freedom. Using the baseline model, the goodness of fit for each distribution is evaluated in term of minimizing the absolute value of the log-likelihood score. As a result, the log-logistic distribution was chosen and became the basis for estimating the regression coefficients of the explanatory variables in the model.

The model was estimated with the LIFEREG procedure in a sequence of six steps by adding explanatory variables into the equation according to the level of analysis (see Table 3): control variables are added first, followed by the population variables, the organization variables, and the individual variables. Model 1, which did not include any explanatory variables, has a log-likelihood score of –1299. The addition of each set of explanatory variables has the effect of improving the log-likelihood score, such that Model 6, which has a score of –295, was chosen as the baseline for comparing other models to understand the effect of each explanatory variable.

The estimation results of Model 6 are provided in Table 3. Among the dummy variables that control for organization type, only the variable that distinguishes research institutes from other types of organizations is significant ($p<.05$ level). Its positive coefficient suggests that researchers employed in research institutes have longer contribution-spans as compared to researchers with other types of affiliations. However, it should be noted that industrial researchers represent only a small segment of the sample (3%); the largest segment being academic (58%).

In addition, the dummy variable that distinguishes between US and non-US researchers is non significant, implying that the contribution-spans of US researchers are no different than their counterparts in other countries. The remaining control variables, the phase and cumulative publication, are both highly significant ($p<.001$ level). The negative sign of the phase variable implies that researchers who were active prior to the field's legitimacy (before 1978) have shorter contribution-spans than those who were active in the growth phase.[25] The coefficient of cumulative publication suggests that the greater the accumulated number of publications in the field, the longer the contribution-spans.

In the case of the population variables, the first- and second-order population terms are significant right from their initial inclusion in Model 3. (The third variable, dispersion, is not significant.) The negative coefficient for population size combined with the positive coefficient for the second-order term implies a U-shaped relationship between population size and researcher contribution-spans (see Figure 4). The data indicate that when the community is small (< 250 researchers), population size is negatively related to contribution-spans and increasingly so until it reaches a size of about 250 individuals; at which point the slope of the curve turns positive. This result suggests that there may be a point of critical

[25]Upon careful reflection, the true effect of this variable is difficult to ascertain given the structure of the present analysis. Understanding the issue of phases of growth in the community might be more appropriately addressed with the use of time-varying covariates.

mass for the community, at which its size is sufficiently large to become significant in increasing contribution-spans. The poplation dispersion variable is also significant ($p<.001$) and has a positive coefficient. Thus, the more disperse researchers are among organizations, the shorter are researcher's contribution-spans.

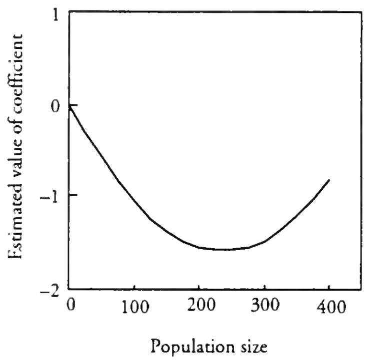

FIGURE 4: *Relationship between population size and researcher contribution-spans*

The second group of explanatory variables, which examine organizational-level factors are added in Model 5, and are both non significant at first. In Model 6, organizational size is weakly significant ($p<.05$) and has a negative coefficient. This suggests that the larger the number of researchers working in cochlear implants in one's organization, the shorter are a researcher's contributions. The cumulative productivity of a researcher's organization is not significant. The relative lack of significance of organization variables is noteworthy given the continued significance of the population variables: it is the size of the entire community, as opposed to a researcher's own organizational effort, which is the contributing factor to the length of contribution-spans.

The last two explanatory variables, the cumulative number of co-authors and an author's cumulative productivity, are highly significant ($p<.001$ level). The data indicate, as might be expected, that researchers who accumulate a greater number of publications will have longer contribution-spans. Perhaps more interesting, is the significance of the cumulative number of individual co-authors a researcher is associated with in the course of their contribution-span. The data show that researchers with a larger network of co-authors will have a longer contribution-span. Since the co-author network can be viewed as a measure of the extent to which a researcher is socially embedded into the research community as a whole, this result

suggests that the degree to which researchers are connected in collaborative activities with their peers in the community is a contributing factor to their longevity in the field.

CONCLUSION

Using the literature as a source of data, this paper provides an analysis of the contribution-spans of researchers in the emerging field of cochlear implants. Non-parametric estimates of the survival rate and hazard rate are made, and it is found that the distribution of contribution-spans follows most closely a log-logistic function. In addition, a statistical model of the relationship between the hazard and a set of explanatory variables is examined.

The findings of this analysis indicate that the sample survival and hazard functions for 1,257 cochlear implant researchers are negatively-sloped and non-monotonic. About 60-percent of the researchers have a contribution-span of more than two years, and the estimated median contribution-span is five years. The risk of a researcher ceasing to contribute to the field is greatest in the first year of their contribution-span. The hazard rate declines sharply after the first year and remains fairly constant until the sixth year, after which it continues to decline rapidly. This result might have an explanation in a behavioral characteristic of the research process. For example, a researcher might be drawn to enter a new field by the prospect of making an important contribution and by the possibility of attracting the resources to underwrite the cost of his or her work. Barring success at either, the researcher might exit the field shortly thereafter for more fertile territory. But if the researcher can initially "survive" in the field, a research program might be established; once in place it would require several years of data collection and analysis. By the sixth year, a critical point is reached, at which time the success of the program leads the researcher to remain in the field with little probability of ever leaving, or lacking success (or intellectual interest), the researcher might decide to move in an alternative research direction.

After controlling for organizational type, geographic location, and the field's phase of growth, it is found that the size of the community, the community's organizational dispersion, the researcher's productivity, and the researcher's collaborative network are statistically significant variables in explaining the length of a researcher's contribution-span. The size of a researcher's organization (in terms of the number of researchers working in the same field) is found to be only weakly significant, and the organization's cumulative productivity in the field is found to be not significant.

These findings provide support to the often-cited importance of the existence of critical mass within a field (that is, the field is sufficiently large in membership to enable a healthy rate of progress to be sustained by researchers). In the case of the cochlear implants community, the data indicate a minimum effective size of about 250 researchers. What is perhaps most interesting, is that critical mass pertains to the community as opposed to the organization: the number of researchers working in the same field within the same organization is less important than the number of researchers working in the field as a whole. The significance of a researcher's co-author network is also interesting in this regard, since it underscores the importance of the researcher being connected to the larger research community in collaborative activities.

The present analysis has certain limitations, some of which are peculiar to literature-based studies and some of which are more generic in nature. Given that the primary interest in this research is to understand the process of a new field's emergence, it will be necessary to construct a data set that implements a time-varying covariate data structure. Such an approach will permit a more careful scrutiny of the dynamic phenomena that affect a researcher's contribution-spans. Furthermore, this approach will allow for the examination of whether or not changes in the hazard rate of a community can serve as an indicator of the future momentum of the field. It is also necessary to determine the extent to which the present and future findings from the cochlear implants field can be generalized to other fields of science and technology and to examine the importance of other explanatory variables in understanding contribution-spans.

Work is currently underway to address these issues. First, a preliminary investigation suggests that data from the literature is structured in such a manner that time-varying covariates should be feasible to create. Second, data sets for ten additional fields are currently being constructed, with fields varying in terms of their size and disciplinary composition, the national and sectoral distribution of their researchers, their commercial impact, and the degree to which they have succeeded in becoming well-established, institutionalized research communities. Third, further studies will be supplemented with other data, derived both from the literature and from other sources. For example, data from the cochlear implant literature is currently being gathered with respect to the nature of the work being conducted by each researcher (such as basic research, applied research, development, or clinical research), and the particular "technological trajectory" each researcher is pursuing (such as a single-channel versus multi-channel device). Other kinds of data, such as annual funding levels and the extent of market commercialization, are being sought as well.

TABLE 1

VARIABLES USED IN THE ANALYSIS AND THEIR DEFINITIONS

CATEGORY	VARIABLE	DESCRIPTION
Dependent variable	Contribution-span	Number of years since researcher's entry in the field
Control variables	*Type of Organization*	
	Academic	Researchers from academic institutions were chosen as the contrast group and were coded as zero
	Industrial	Researchers employed in industry = 1; otherwise = 0
	Research	Researchers employed in private research institutes = 1; otherwise = 0
	Government	Researchers employed in government institutes = 1; otherwise = 0
	Geographic location	
	USA	Researchers employed in the U.S. = 1; otherwise = 0
	Phase of growth	
	Pre-legimacy phase	Before 1978 = 1; 1978 and thereafter=0
	Cum publication	Cumulative number of publications in the field
Population variables	Population size	Total number of authors in the field
	Population^2/1000	Second order term of population size
	Population dispersion	Measure of researcher dispersion among organizations (Hirfindahl index)
Organization variables	Organization size	Number of researchers working in organization in the field
	Cum organization productivity	Cumulative number of publications by organization in the field
Individual variables	Cumulative co-authors	Cumulative number of individual co-authors affilated with the researcher
	Cumulative author productivity	Cumulative number of publications in the field by researcher

TABLE 2

ML ESTIMATION OF CONTRIBUTION-SPANS USING DIFFERENT DISTRIBUTIONS

	Exponential	Weibull	Gamma	Log-logistic
Intercept	0.750 (0.559)	0.297 (0.219)	−0.062*** (0.168)	0.081 (0.172)
Type of Organization				
Industrial	−0.248 (0.260)	−0.203** (0.103)	−0.037 (0.089)	−0.109 (0.081)
Research	0.202* (0.101)	0.255*** (0.040)	0.072 (0.038)	0.076* (0.033)
Government	0.047 (0.585)	−0.070 (0.230)	0.006 (0.236)	0.071 (0.168)
Geographic location				
USA	−0.012 (0.091)	0.021 (0.036)	0.002 (0.034)	−0.000 (0.029)
Phase of growth				
Pre-legitimacy phase	−2.245*** (0.416)	−1.046*** (0.165)	−0.937*** (0.152)	−0.984*** (0.140)
Cumulative publications in field	0.004*** (0.000)	0.002*** (0.000)	0.002*** (0.000)	0.002*** (0.000)
Population size	−0.036*** (0.004)	−0.015*** (0.001)	−0.008*** (0.001)	−0.013*** (0.001)
$\text{Population}^2/1000$	0.080*** (0.010)	0.032*** (0.004)	0.015*** (0.003)	0.028*** (0.003)
Population dispersion	0.015** (0.005)	−0.004** (0.002)	0.009*** (0.001)	0.005*** (0.002)
Organization size	−0.008 (0.012)	−0.006 (0.004)	−0.012* (0.005)	−0.008* (0.004)
Cum organization productivity	0.001 (0.004)	0.000 (0.002)	0.003* (0.001)	0.002 (0.001)
Cumulative co-authors	0.023 (0.020)	0.012 (0.008)	0.038*** (0.008)	0.023*** (0.007)
Cumulative author productivity	0.315*** (0.038)	0.393*** (0.021)	0.171*** (0.009)	0.307*** (0.014)
Scale parameter		0.393 (0.012)	0.404 (0.000)	0.199 (0.007)
Shape parameter			−0.495 (0.000)	
Log-Likelihood	−689	−400	−365	−295

NOTE: 1.) Total number of researchers = 1257
2.) Figures in parentheses are standard errors of estimates
3.) Significance level: *< .05; **< .01 ***< .001

TABLE 3

ML ESTIMATION OF CONTRIBUTION-SPANS:
LOG-LOGISTIC DISTRIBUTION

	Model1	Model2	Model3	Model4	Model5	Model6
Intercept	0.964***	–0.197***	–0.749***	0.195	0.087	0.081
	(0.039)	(0.024)	(0.037)	(0.145)	(0.173)	(0.172)
Cum author productivity		0.412***	0.358***	0.340***	0.329***	0.307***
		(0.012)	(0.012)	(0.012)	(0.012)	(0.014)
Type of Organization						
Industrial			–0.117	–0.094	–0.102	–0.109
			(0.083)	(0.079)	(0.083)	(0.081)
Research			0.060	0.094**	0.091**	0.076*
			(0.034)	(0.031)	(0.069)	(0.033)
Government			–0.142	0.062	0.040	0.171
			(0.178)	(0.160)	(0.168)	(0.168)
Geographic location						
USA			0.035	0.007	–0.011	–0.000
			(0.029)	(0.027)	(0.029)	(0.029)
Phase of growth						
Pre-legitimacy phase			0.309***	–0.881***	–0.994***	–0.984***
			(0.053)	(0.120)	(0.142)	(0.140)
Cum publication			0.001***	0.001***	0.002***	0.002***
			(0.000)	(0.000)	(0.000)	(0.000)
Population size				–0.014***	–0.014***	–0.013***
				(0.002)	(0.001)	(0.001)
Population^2				0.029***	0.029***	0.028***
				(0.002)	(0.003)	(0.003)
Population dispersion				0.004**	0.006***	0.005***
				(0.001)	(0.002)	(0.002)
Organization size					–0.004	–0.008*
					(0.004)	(0.004)
Cum org productivity					0.001	0.002
					(0.001)	(0.001)
Cum co-authors						0.023***
						(0.007)
Scale parameter	0.633	0.241	0.216	0.192	0.201	0.199
	(0.021)	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)
Log-Likelihood	–1299	–705	–436	–303	–301	–295

NOTE: 1.) Total number of researchers = 1257 (604 non-censored)
2.) Figures in parentheses are standard errors of estimates
3.) Significance level: *< .05; **< .01; *** < .001

Date Due 9-8 92